Tekstherkenning: ScanSoft OmniPage Pro12

Leer je pc lezen

Het volstaat allang niet meer als een *OCR*-pakket gewone tekst foutloos kan scannen, lezen, begrijpen, en in een Word-document plakken. De documenten die je wil scannen worden steeds complexer van opmaak, en daar moeten tekstherkenningspakketten ook mee uit de voeten kunnen. ScanSoft sleutelt daarom voortdurend aan zijn OCR-toepassingen. Dat moet je nu merken aan de nieuwste OmniPage Pro 12.

OCR-software moet in de eerste plaats gescande stukken tekst omzetten naar een tekst die bewerkbaar en doorzoekbaar is op je pc. De nauwkeurigheid van de vorige OmniPage Pro kon voor het herkennen van gewone standaardtekst niet veel beter. Toch zijn er in versie 12 heel wat nieuwigheden te melden. Vooral complexere documenten en *pdf*-bestanden converteren was toch nog voor verbetering vatbaar.



Hardop lezen

Eén van de belangrijkste verbeteringen is het herkennen en converteren van pdf-bestanden. De nieuwe OmniPage 12 bevat immers een technologie die pdf-documenten vergelijkt in verschillende lagen. Daardoor converteert het programma zowel pdf-bestanden die volledig uit tekst bestaan, als pdf-bestanden met afbeeldingen en tekst, ongeveer 80 % nauwkeuriger. De 'bewerkte pdf'-uitvoer is een nieuwe uitvoermodus waarmee je tekst en opmaak van een document ingrijpend kan wijzigen in de OmniPage proeflezer. Het resultaat is een goed opgemaakt en doorzoekbaar pdf-bestand. In het pakket zitten verschillende OCR-engines ingebakken, en die zorgen samen voor een nog nauwkeuriger eindresultaat als je een complexe lay-out moet bewaren. OmniPage herkent nu documenten in 114 talen. De nauwkeurigheid kon ook opgetrokken worden omdat de software nu z'n eigen leesvaardigheid oefent op individuele letters en symbolen, bij wijze van aanvulling op de reeds bestaande IntelliTrain proefleesfunctie. Dokters en advokaten zullen de nieuwe ingebouwde woordenlijst met medische en juridische terminologie kunnen appreciëren, want ook die wordt nu sneller herkend.

Tot de meer geavanceerde nieuwigheden behoort de mogelijkheid om twee pagina's tegelijk te scannen. Dat is handig wanneer je boeken en artikels uit tijdschriften wil scannen. OmniPage scheidt de pagina's netjes, en elke pagina afzonderlijk wordt, los van de ander, 'afgetast' en herkend. OmniPage 12 werkt nog beter samen met Office, zowel Office 2000 als Office XP. Office XP heeft overigens zelf een beperkte OCR-functie aan boord, maar gelukkig vormt die geen belemmering voor OmniPage. Het pakket nestelt zich in de menu'tjes van Office, en zo kan je vanuit Word, Excel en andere Office-toepassingen rechtstreeks een OCR-opdracht geven. Het trucje werkt nu ook vanuit WordPerfect 10.
Als toetje kan je met de RealSpeak-voorleesfunctie je gescande en herkende teksten laten voorlezen. Deze gimmick komt rechtstreeks overwaaien uit Ieper en is een restant van Lernout & Hauspie, waarvan ScanSoft een groot deel heeft overgenomen.

Twee smaken

ScanSoft OmniPage Pro 12 is beschikbaar in twee versies. De standaardversie heeft alle bovenstaande vernieuwingen en zal wel het meest gebruikt worden. De veel duurdere OmniPage Pro 12 Office versie heeft nog een aantal extra snufjes meegekregen, die vooral bedoeld zijn voor bedrijven. Zo kan die versie bijvoorbeeld documenten bewaren in *XML*, heeft ze netwerkondersteuning voor netwerkscanners en kan ze grotere scanopdrachten automatisch uitvoeren. Verder ondersteunt deze versie Open eBook en Tablet PC's.

— Dirk Peeters —

CONCLUSIE

ScanSoft heeft aan OmniPage Pro 12 heel wat nuttige verbeteringen aangebracht, die het product vooral voor complexere documenten flink wat nauwkeuriger maken. Het spreekt dat al dat fraais ook een prijskaartje heeft, en enkel zijn geld waard is als je geregeld papieren documenten in handen krijgt zonder een digitale tegenhanger.

PRODUCTINFO
Product: OmniPage Pro 12
Producent: ScanSoft
Prijs: € 175 (Office-versie kost € 757 of € 288 voor een upgrade)
Systeemvereisten: Pentium of hoger, Windows 98 of hoger (incl. XP), 64 MB RAM-geheugen (128 MB aanbevolen), 140 MB vrije ruimte op de harde schijf, SVGA-kaart en monitor, TWAIN-scanner.
Url: [www.omnipage.com]

VAKTAAL

OCR: Optical Character Recognition. Een technologie die je in staat stelt teksten in te scannen en die daarna ook te bewerken. Zo vermijd je dat je een tekst helemaal moet overtypen.

Pdf: Portable Document File. Een pdf lees je met behulp van de gratis Adobe Acrobat Reader, en behoudt altijd z'n zelfde lay-out.

SVGA-kaart: Een beeldstandaard voor de grafische kaart.

TWAIN: Technology Without An Interesting Name. Software waarmee scanners en digitale camera's hun beelden kunnen doorgeven aan grafische programma's.

Xml: Een programmeertaal om webpagina's te maken.